LAPORAN UJIAN AKHIR SEMESTER

STATISTIKA DESKRIPTIF

CLUSTERING

NAMA : **MUKHAMAD IKHSANUDIN**

NIM : **082011633086**

DOSEN PENGAMPU : **Drs. ETO WURYANTO, DEA.**

196609281991021001

PROGRAM STUDI S1 SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS AIRLANGGA

2021

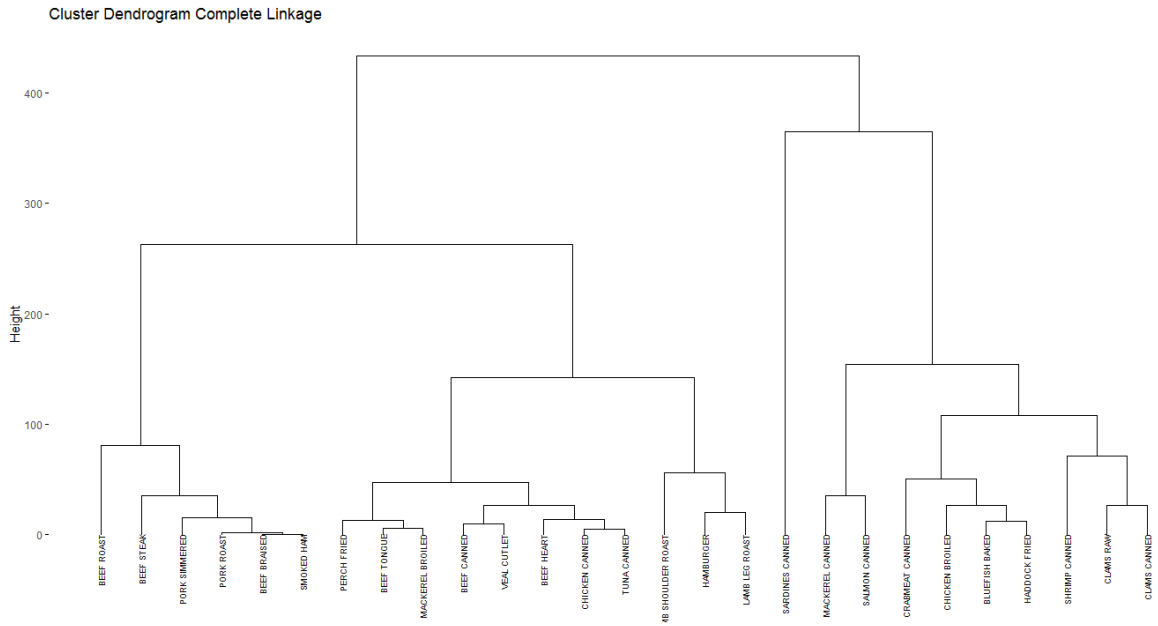
```{R}
# Clustering
library(flexclust)
library(cluster)
library(factoextra)
library(mclust)
library(Gmedian)


# 1. Dataset Preparation
data("nutrient")
DataClust <- nutrient
Head(DataClust)
```

Description: df[,5] [6 x 5]

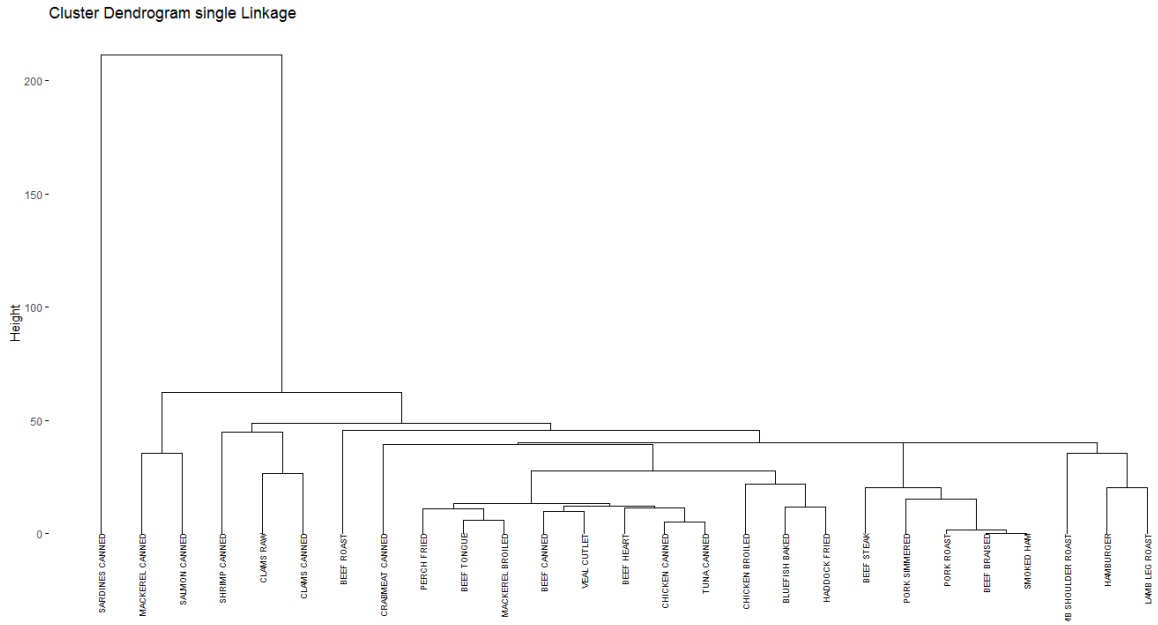
	energy <int>	protein <int>	fat <int>	calcium <int>	iron <dbl>
BEEF BRAISED	340	20	28	9	2.6
HAMBURGER	245	21	17	9	2.7
BEEF ROAST	420	15	39	7	2.0
BEEF STEAK	375	19	32	9	2.6
BEEF CANNED	180	22	10	17	3.7
CHICKEN BROILED	115	20	3	8	1.4

6 rows

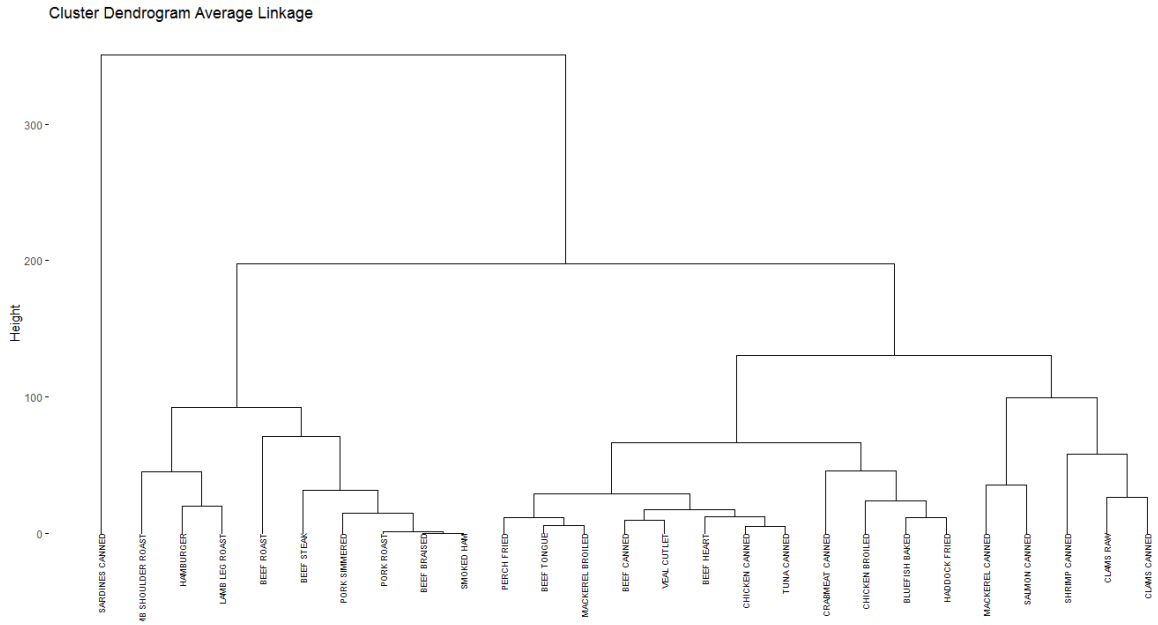
```
# Metode Agglomerative
# Complete Linkage
Clust_Com <- hclust(dist(DataClust), method = "complete")
Clust_Com
fviz_dend(Clust_Com, cex = 0.4, main = "Cluster Dendrogram Complete
Linkage")
```



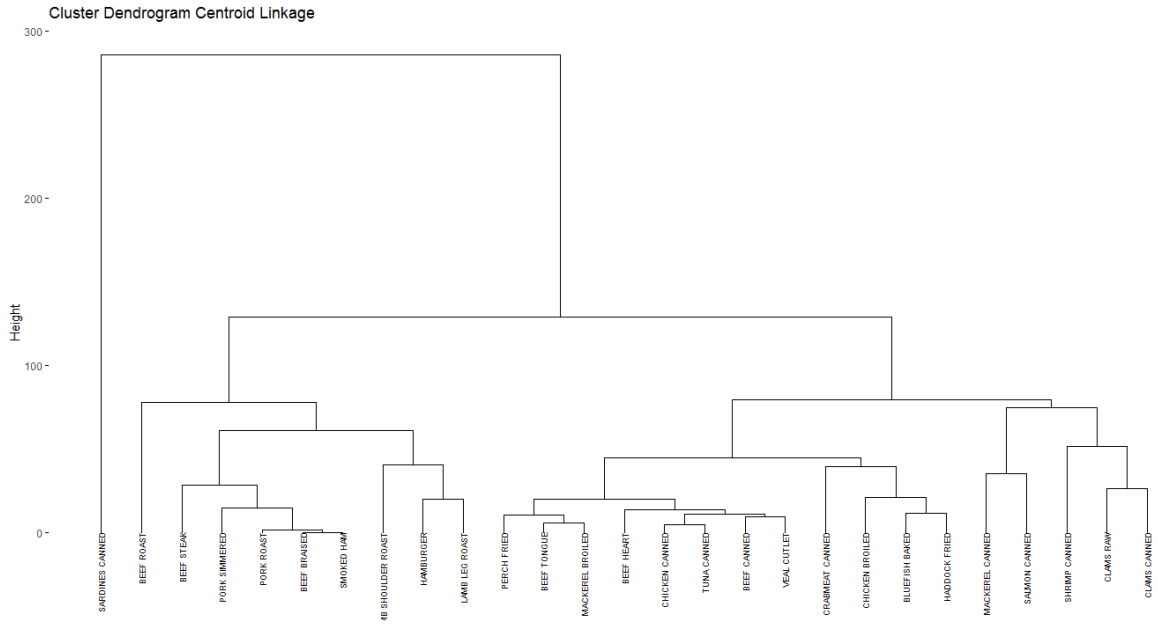
```
# Single Linkage
Clust_Sin <- hclust(dist(DataClust), method = "single")
Clust_Sin
fviz_dend(Clust_Sin, cex = 0.4, main = "Cluster Dendrogram single
Linkage")
```



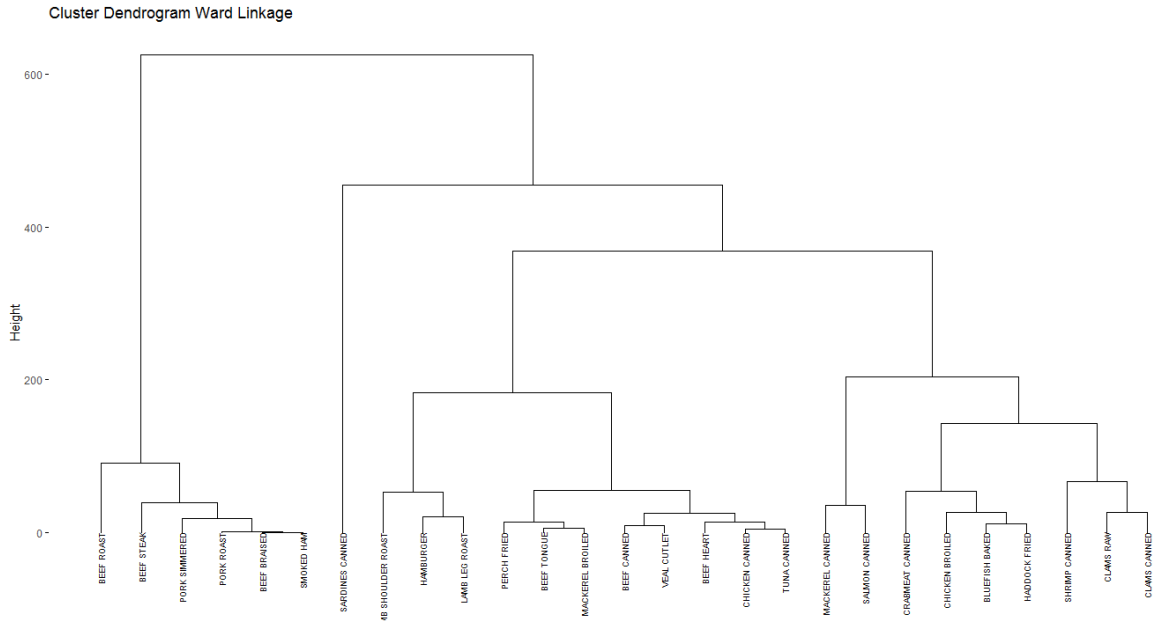

```
# Average Linkage
Clust_Ave <- hclust(dist(DataClust), method = "average")
Clust_Ave
fviz_dend(Clust_Ave, cex = 0.4, main = "Cluster Dendrogram Average
Linkage")
```




```
# Centroid Linkage
Clust_Cen <- hclust(dist(DataClust), method = "centroid")
Clust_Cen
fviz_dend(Clust_Cen, cex = 0.4, main = "Cluster Dendrogram Centroid
Linkage")
```

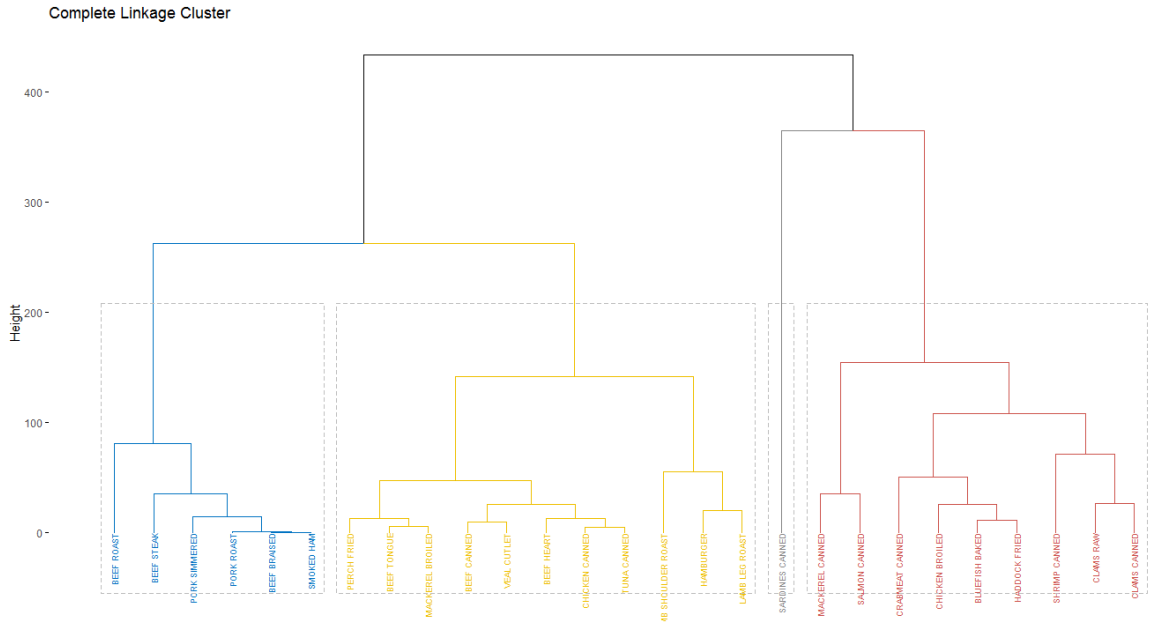



```
# Ward Linkage
Clust_War <- hclust(dist(DataClust), method = "ward.D2")
Clust_War
fviz_dend(Clust_War, cex = 0.4, main = "Cluster Dendrogram Ward
Linkage")
```



Berdasarkan 5 plot dendogram di atas dengan menggunakan metode aglomerative yang terdiri antara metode centroid, single-linkage, complete-linkage, average-linkage dan ward serta terdapat perhitungan matriks distance dengan fungsi dist(dataset), didapatkan informasi bahwa clustering untuk dataset *nutrient* dapat terbagi menjadi 4 klaster utama. Matriks distance merupakan matriks yang berisikan pasangan jarak antar objek yang dihitung dengan Euclidian Distance.

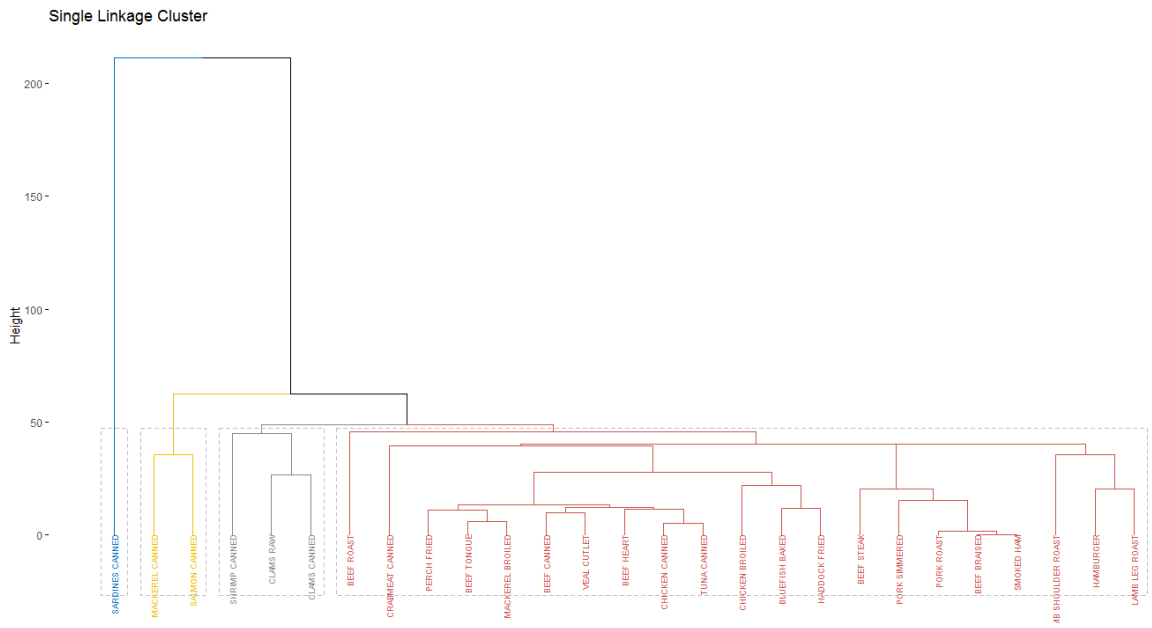
```
# Banyaknya Clustering Berdasarkan Syntax dan Dendogram
# Complete Linkage
fviz_dend(Clust_Com, k = 4, k_colors = "jco", rect = TRUE,
          cex = 0.4, main = "Complete Linkage Cluster")
```



Berdasarkan plot dendogram dengan metode complete di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

- Kluster 1 merupakan kluster dengan jumlah energi (dalam kalori) terbanyak
  beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham
- Kluster 2 merupakan kluster dengan jumlah energi menengah ke atas
  perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, lamb shoulder roast, hamburger, dan lamb leg roast
- Kluster 3 merupakan kluister dengan jumlah energi menengah ke bawah
  sardines canned
- Kluster 4 merupakan kluster dengan jumlah energy terkecil
  mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned

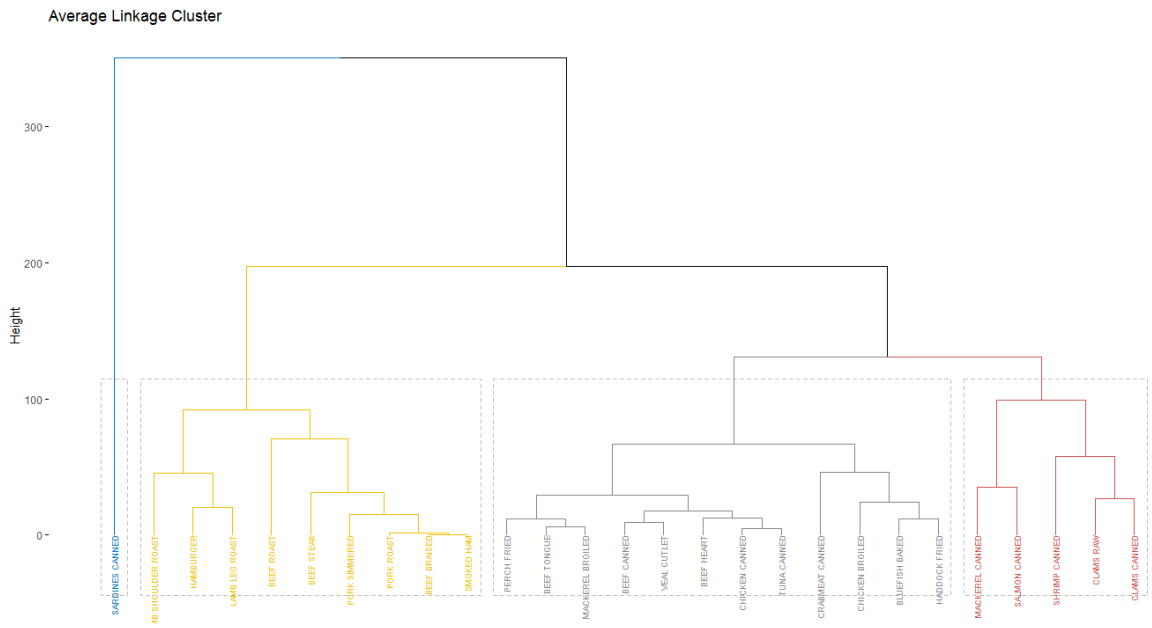
```
# Single Linkage
fviz_dend(Clust_Sin, k = 4, k_colors = "jco", rect = TRUE,
          cex = 0.4, main = "Single Linkage Cluster")
```



Berdasarkan plot dendogram dengan metode single di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

- Kluster 1 merupakan kluster dengan jumlah kalsium (mg) terbanyak
  sardines canned
- Kluster 2 merupakan kluster dengan jumlah kalsium (mg) menengah ke atas
  mackerel canned dan salmon canned
- Kluster 3 merupakan kluster dengan jumlah kalsium (mg) menengah ke bawah
  shrimp canned, clams raw, dan clams canned
- Kluster 4 merupakan kluster dengan jumlah kalsium (mg) terkecil
  beef roast, crabmeat canned, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, chicken broiled, bluefish baked, haddock fried, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast

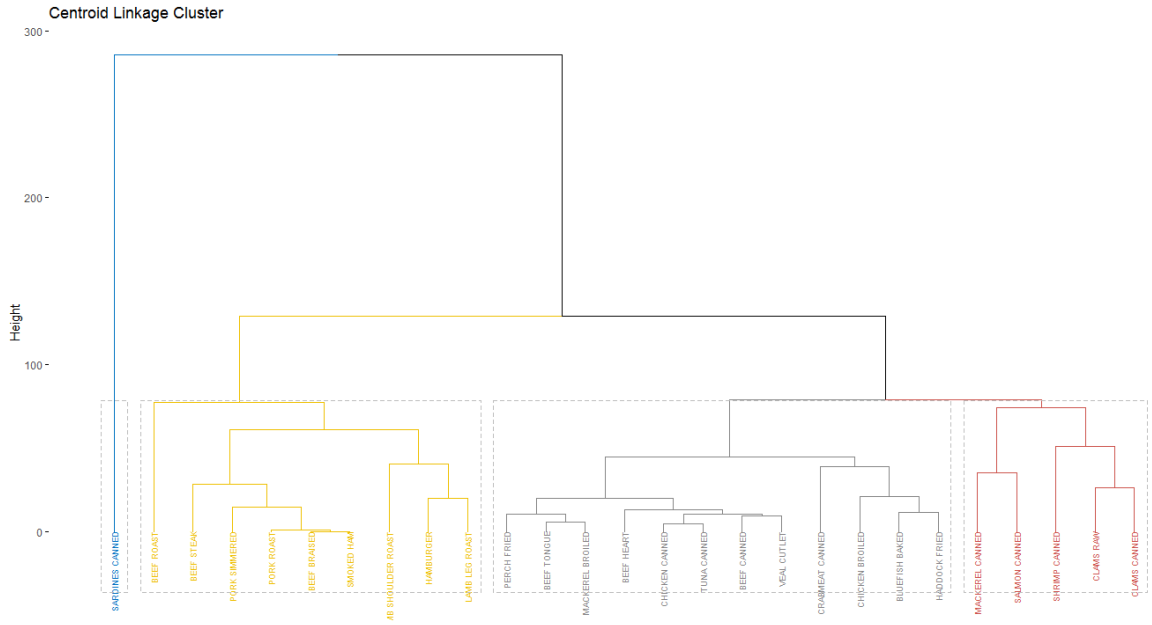
```
# Average Linkage
fviz_dend(Clust_Ave, k = 4, k_colors = "jco", rect = TRUE,
          cex = 0.4, main = "Average Linkage Cluster")
```



Berdasarkan plot dendogram dengan metode average di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

- Kluster 1 merupakan kluster dengan rata-rata nutrisi terbanyak
  sardines canned
- Kluster 2 merupakan kluster dengan rata-rata nutrisi menengah ke atas
  lamb shoulder roast, hamburger, lamb leg roast, beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham
- Kluster 3 merupakan kluster dengan rata-rata nutrisi menengah ke bawah
  perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried
- Kluster 4 merupakan kluster dengan rata-rata nutrisi terkecil
  mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned

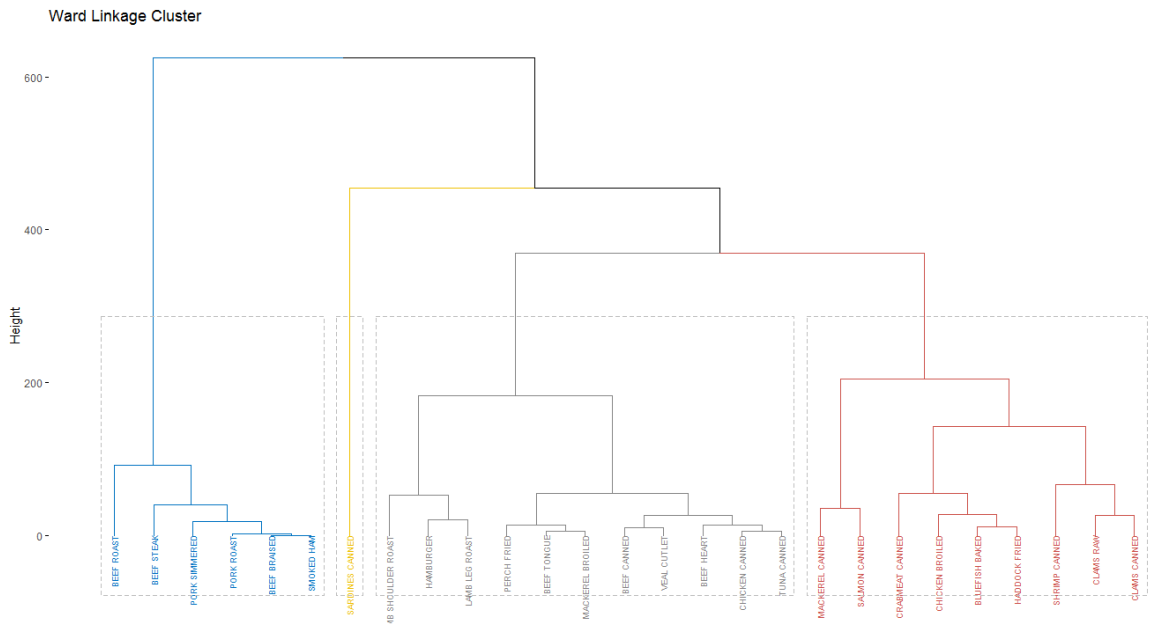
```
# Centroid Linkage
fviz_dend(Clust_Cen, k = 4, k_colors = "jco", rect = TRUE,
          cex = 0.4, main = "Centroid Linkage Cluster")
```



Berdasarkan plot dendogram dengan metode centroid di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

- Kluster 1 merupakan kluster dengan jarak menuju nilai pusat terdekat
  sardines canned
- Kluster 2 merupakan kluster dengan jarak menuju nilai pusat menengah dekat
  beef roast, beef steak, pork simmered, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast
- Kluster 3 merupakan kluster dengan jarak menuju nilai pusat menengah jauh
  perch fried, beef tongue, mackerel broiled, beef heart, chicken canned, tuna canned, beef canned, veal cutlet, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried
- Kluster 4 merupakan kluster dengan jarak menuju nilai pusat terjauh
  mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned

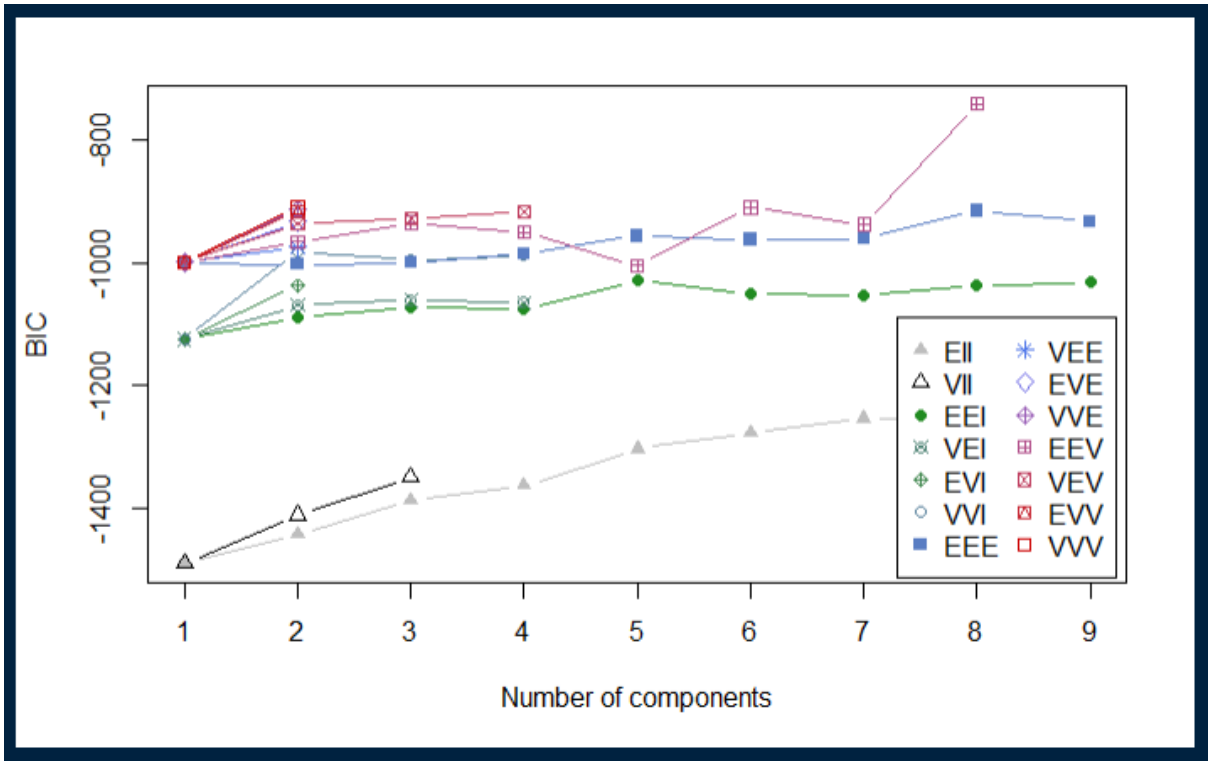
```
# Ward Linkage
fviz_dend(Clust_War, k = 4, k_colors = "jco", rect = TRUE,
          cex = 0.4, main = "Ward Linkage Cluster")
```



Berdasarkan plot dendogram dengan metode ward di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

- Kluster 1 merupakan kluster dengan jumlah beberapa variabel terbanyak
  beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham
- Kluster 2 merupakan kluster dengan jumlah beberapa variabel menengah ke atas
  sardines canned
- Kluster 3 merupakan kluster dengan jumlah beberapa variabel menengah ke bawah
  lamb shoulder roast, hamburger, lamb leg roast, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, dan tuna canned
- Kluster 4 merupakan kluster dengan jumlah beberapa variabel terkecil
  mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned

```
# Banyaknya Clustering Berdasarkan BIC
New_DataClust <- as.matrix(nutrient)
New_DataClust <- Mclust(nutrient)
plot(New_DataClust)
```

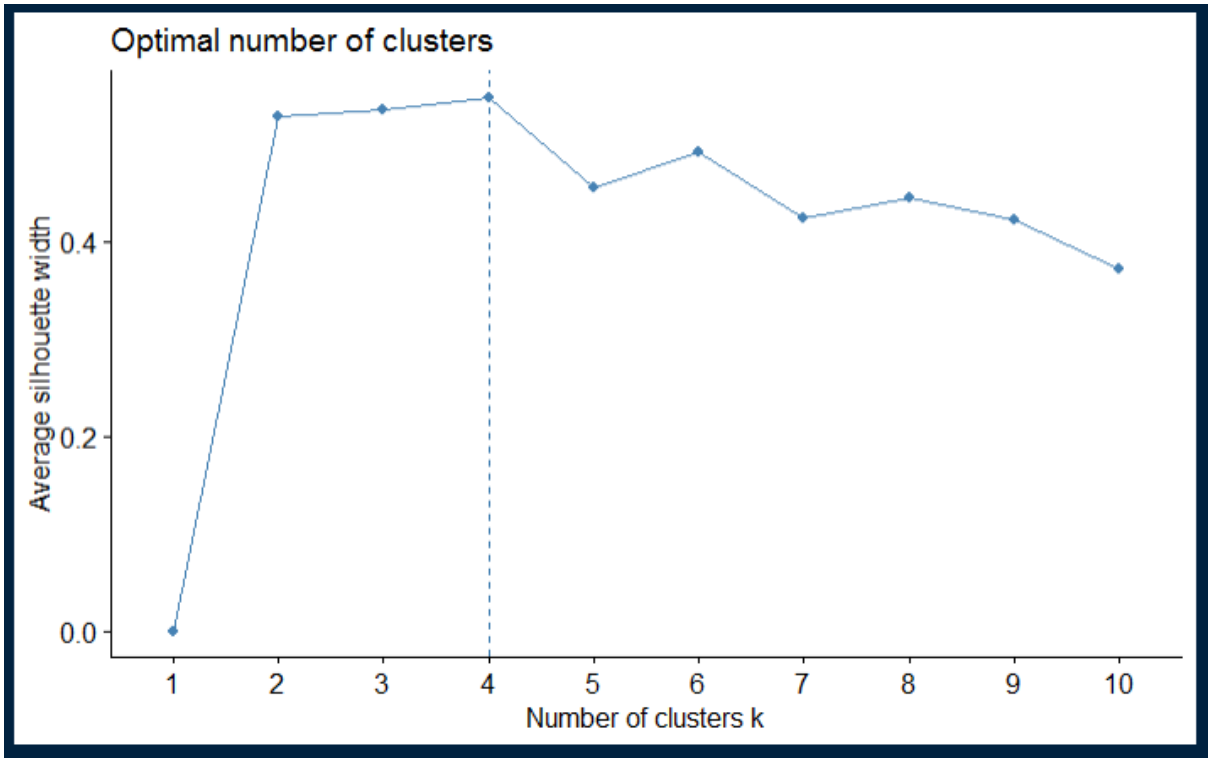


Berdasarkan dari plot BIC tersebut di dapatkan bahwa data *nutrient* dapat dibagi menjadi 1 dan 4. Jika dilihat berdasarkan plot nampak bahwa clustering terbaik yaitu dibagi menjadi 4 kluster. Hal ini ditunjukkan dengan semakin banyak komponen yang ada pada suatu *number of components* maka dapat disimpulkan bahwa *number* tersebut bisa menjadi *clustering* terbaik.

No	Metode	Nomer klaster	Anggota klaster
1	Complete-Linkage	1	beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham
		2	perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, lamb shoulder roast, hamburger, dan lamb leg roast
		3	sardines canned
		4	mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned
2	Single-Linkage	1	sardines canned
		2	mackerel canned dan salmon canned
		3	shrimp canned, clams raw, dan clams canned

		4	beef roast, crabmeat canned, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, chicken broiled, bluefish baked, haddock fried, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast
3	Average-Linkage	1	sardines canned
		2	lamb shoulder roast, hamburger, lamb leg roast, beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham
		3	perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried
		4	mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned
4	Centroid	1	sardines canned
		2	beef roast, beef steak, pork simmered, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast
		3	perch fried, beef tongue, mackerel broiled, beef heart, chicken canned, tuna canned, beef canned, veal cutlet, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried
		4	mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned
5	Ward	1	beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham
		2	sardines canned
		3	lamb shoulder roast, hamburger, lamb leg roast, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, dan tuna canned
		4	mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned

```
# Nilai K-Means dan Clustering
fviz_nbclust(nutrient, kmeans, method = "silhouette")
```

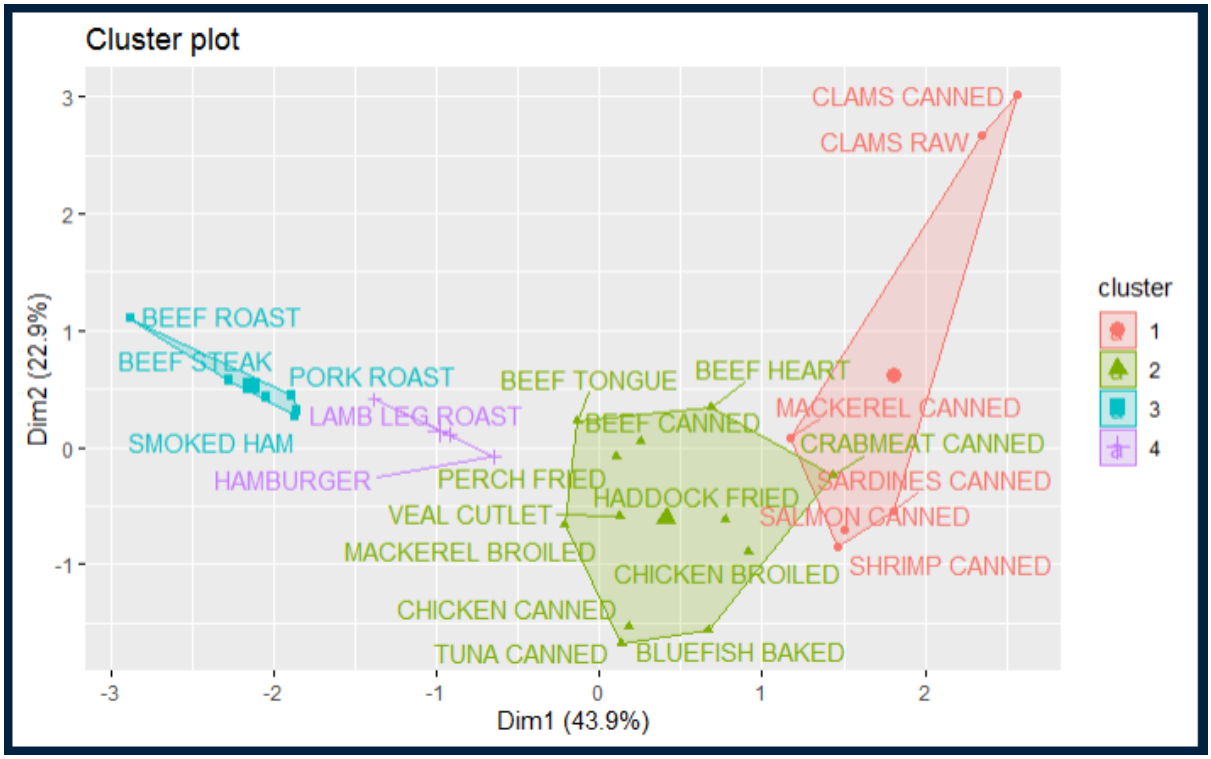


Pada visualisasi dengan metode silhouette, nilai K optimum dapat dilihat dengan titik tertinggi yang ada pada grafik. Namun, bisa juga dilihat dengan titik kedua tertinggi yang ada pada grafik sehingga nilai K optimum pada metode ini adalah antara 3 dan 4. Sehingga dari kedua metode yang digunakan dapat disimpulkan bahwa nilai k optimum adalah 4.

```
Data_kmean <- kmeans(nutrient, 2)
Data_kmean
```

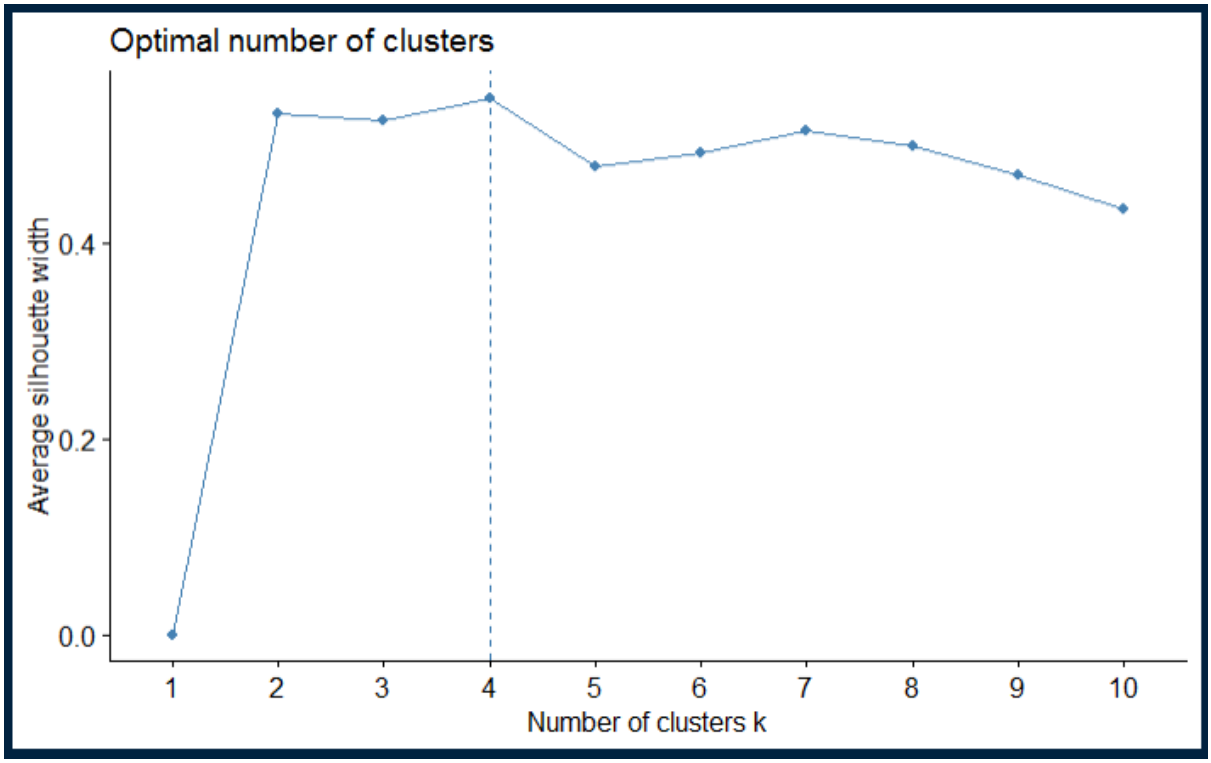
```
K-means clustering with 4 clusters of sizes 6, 12, 6, 3

Cluster means:
    energy  protein       fat    calcium     iron
1 113.3333 16.00000  4.333333 156.166667 3.166667
2 161.6667 20.50000  7.500000  14.250000 1.925000
3 361.6667 18.66667 31.000000   8.666667 2.433333
4 270.0000 19.66667 20.666667   9.000000 2.533333

Clustering vector:
       BEEF BRAISED           HAMBURGER         BEEF ROAST          BEEF STEAK        BEEF CANNED     CHICKEN BROILED
                  3                   4                   3                   3                  2                   2
     CHICKEN CANNED         BEEF HEART    LAMB LEG ROAST LAMB SHOULDER ROAST        SMOKED HAM         PORK ROAST
                  2                   2                   4                   4                  3                   3
      PORK SIMMERED        BEEF TONGUE        VEAL CUTLET      BLUEFISH BAKED         CLAMS RAW       CLAMS CANNED
                  3                   2                   2                   2                  1                   1
    CRABMEAT CANNED       HADDOCK FRIED    MACKEREL BROILED     MACKEREL CANNED       PERCH FRIED     SALMON CANNED
                  2                   2                   2                   1                  2                   1
    SARDINES CANNED         TUNA CANNED       SHRIMP CANNED
                  1                   2                   1

Within cluster sum of squares by cluster:
[1] 73169.233 15536.079  5142.253  1587.420
 (between_SS / total_SS =  77.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

```
fviz_cluster(Data_kmean, data = nutrient, rep = TRUE)
```



Berdasarkan cluster plot di atas, data *nutrient* dikelompokkan menjadi 4 kluster. Persebaran data dari dimensi 1 dan 2 sebesar 66,8%. Dari plot tersebut, kluster yang tertinggi adalah kluster 1 dengan data clams canned dan clams raw. Lalu kluster 2 berada di bawahnya. Namun beberapa data pada kluster 2 memiliki nilai dimensi 2 negatif seperti tuna canned dan bluefish baked. Kluster 3 hanya memiliki 2 data yaitu lamb leg roast dan hamburger. Semua data kluster 4 berada pada dimensi 1 negatif.

```
# Nilai K-Medoids dan Clusteirng
fviz_nbclust(nutrient, pam, method = "silhouette")
```

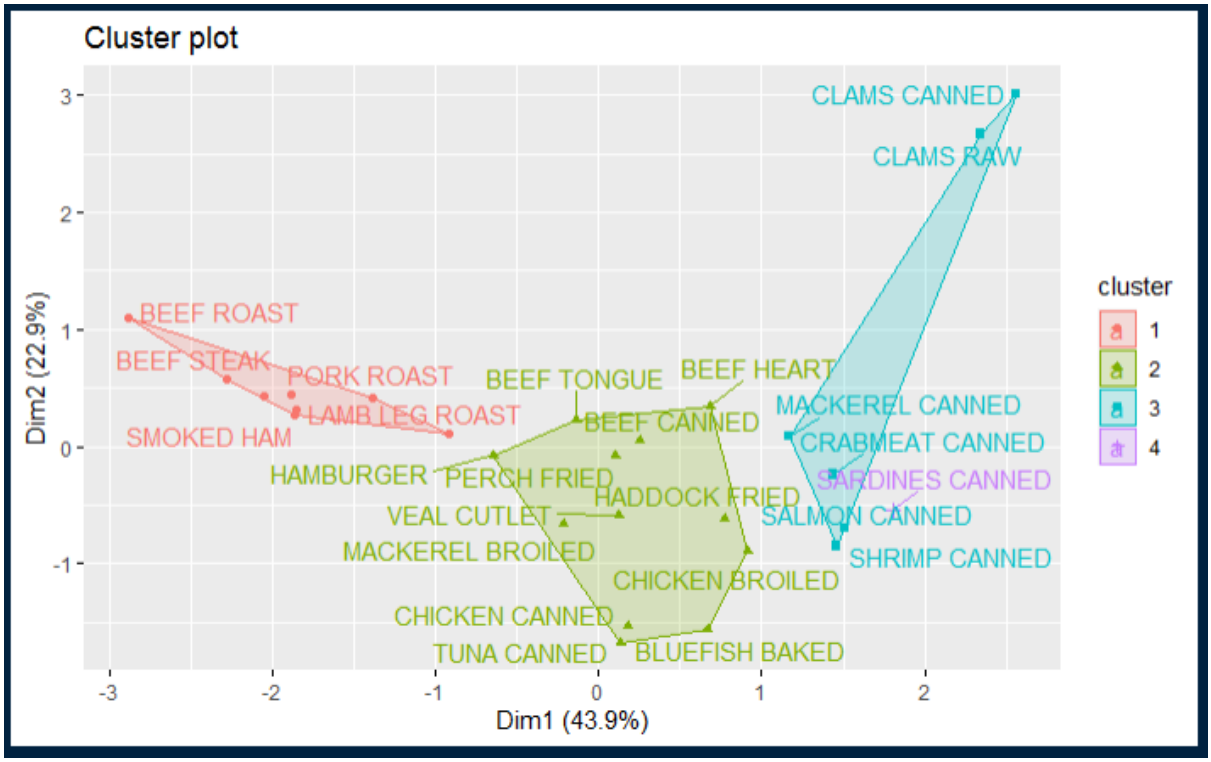


Pada visualisasi dengan metode silhouette, nilai K optimum dapat dilihat dengan titik tertinggi yang ada pada grafik. Namun, bisa juga dilihat dengan titik kedua tertinggi yang ada pada grafik sehingga nilai K optimum pada metode ini adalah antara 3 dan 4. Sehingga dari kedua metode yang digunakan dapat disimpulkan bahwa nilai k optimum adalah 4.

```
Data_kmedoid <- pam(nutrient, 2)
summary(Data_kmedoid)
Data_kmedoid
```

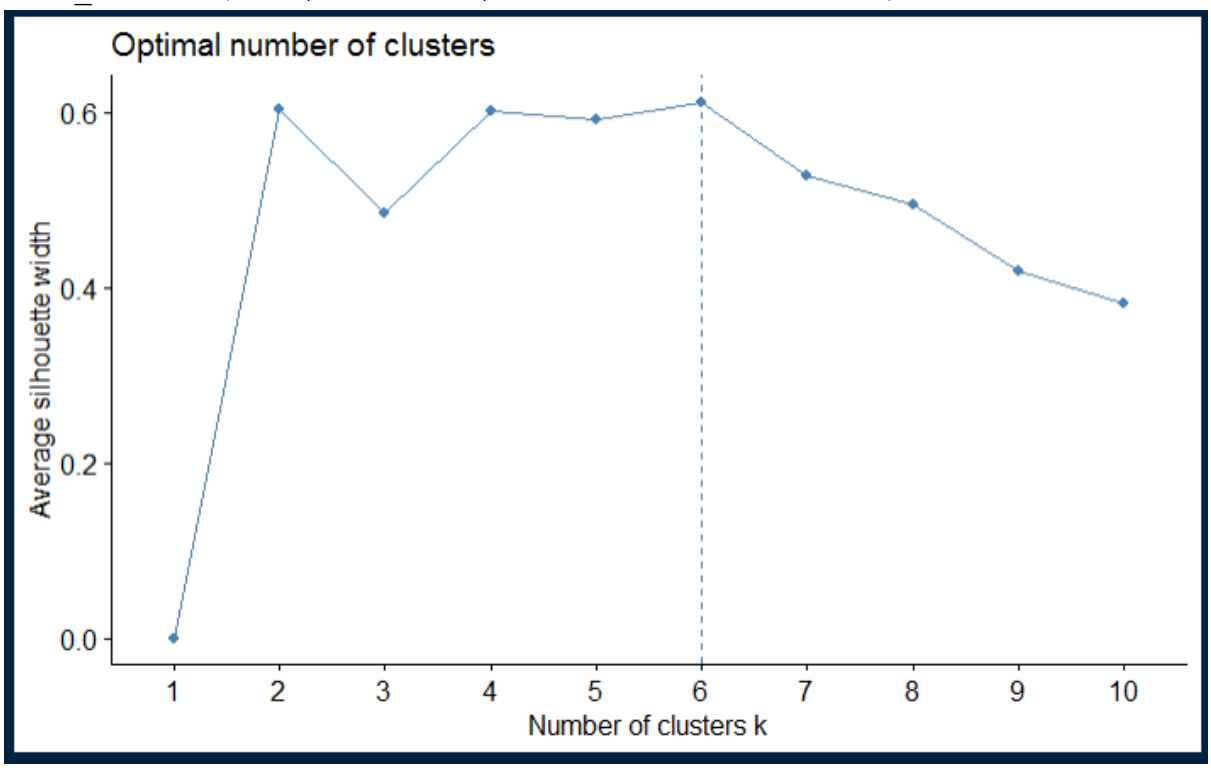
```
Medoids:
               ID energy protein fat calcium iron
SMOKED HAM     11    340      20  28       9  2.5
CHICKEN CANNED  7    170      25   7      12  1.5
SHRIMP CANNED  27    110      23   1      98  2.6
SARDINES CANNED 25   180      22   9     367  2.5
Clustering vector:
       BEEF BRAISED           HAMBURGER         BEEF ROAST          BEEF STEAK         BEEF CANNED
                  1                   2                  1                   1                   2
    CHICKEN BROILED      CHICKEN CANNED         BEEF HEART     LAMB LEG ROAST LAMB SHOULDER ROAST
                  2                   2                  2                   1                   1
         SMOKED HAM          PORK ROAST       PORK SIMMERED         BEEF TONGUE         VEAL CUTLET
                  1                   1                  1                   2                   2
     BLUEFISH BAKED           CLAMS RAW       CLAMS CANNED     CRABMEAT CANNED       HADDOCK FRIED
                  2                   3                  3                   3                   2
   MACKEREL BROILED     MACKEREL CANNED         PERCH FRIED      SALMON CANNED     SARDINES CANNED
                  2                   3                  2                   3                   4
        TUNA CANNED       SHRIMP CANNED
                  2                   3
Objective function:
   build     swap
33.91351 33.72627

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"  "clusinfo"   "silinfo"    "diss"
 [9] "call"       "data"
```

```
fviz_cluster(Data_kmedoid, data = nutrient, rep = TRUE)
```



Berdasarkan cluster plot di atas, data *nutrient* dikelompokkan menjadi 4 kluster. Persebaran data dari dimensi 1 dan 2 sebesar 66,8%. Dari plot tersebut, kluster yang tertinggi adalah kluster 3 dengan data clams canned dan clams raw. Lalu kluster 2 berada di bawahnya. Namun beberapa data pada kluster 2 memiliki nilai dimensi 2 negatif seperti tuna canned dan bluefish baked. Kluster 1 hanya memiliki 1 data yaitu sardines caneed dengan nilai dimensi 2 negatif. Semua data kluster 4 berada pada dimensi 1 negatif.

```
# Nilai KMedians
fviz_nbclust(milk, kGmedian, method = "silhouette")
```



Pada visualisasi dengan metode silhouette, nilai K optimum dapat dilihat dengan titik tertinggi yang ada pada grafik. Namun, bisa juga dilihat dengan titik kedua tertinggi yang ada pada grafik sehingga nilai K optimum pada metode ini adalah antara 2, 4, dan 6. Sehingga dari kedua metode yang digunakan dapat disimpulkan bahwa nilai k optimum adalah 6.

```
Data_kmedian <- kGmedian(nutrient, 6)
Data_kmedian
```

```
$cluster
      [,1]
 [1,]    1
 [2,]    3
 [3,]    1
 [4,]    1
 [5,]    2
 [6,]    2
 [7,]    2
 [8,]    2
 [9,]    3
[10,]    1
[11,]    1
[12,]    1
[13,]    1
[14,]    3
[15,]    2
[16,]    2
[17,]    6
[18,]    6
[19,]    6
[20,]    2
[21,]    3
[22,]    4
[23,]    3
[24,]    4
[25,]    5
[26,]    2
[27,]    6
```

```
$centers
     energy  protein       fat    calcium     iron
1 352.79870 18.82487 29.979704   8.742562 2.454090
2 159.52624 22.72529  6.632552  13.653360 2.680313
3 228.31705 19.27179 15.657911   8.709502 2.192537
4 140.21442 16.42245  7.310219 157.844890 1.335310
5 180.00000 22.00000  9.000000 367.000000 2.500000
6  74.32389 12.56184  1.184858  74.585093 4.347628

$withinsrs
          [,1]
[1,] 21.757250
[2,] 23.920988
[3,] 17.544362
[4,]  5.603242
[5,]  0.000000
[6,] 14.311382

$size
     [,1]
[1,]    7
[2,]    8
[3,]    5
[4,]    2
[5,]    1
[6,]    4
```

Kesimpulan

Dataset *nutrient* sangat optimal apabila dikelompokkan menjadi 4 kluster, baik menggunakan metode complete-linkage, single-linkage, average-linkage, centroid, maupun ward. Masing-masing metode memiliki pembagian data yang belum tentu sama dengan metode yang lain.